IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re Patent Application of	Confirmation No.: 6586
ASSADIAN et al	Atty. Ref.: 36-1982
Serial No. 10/573,152	TC/A.U.: 2129
Filed: March 23, 2006	Examiner: K. Bharadwaj
For: INFORMATION RETRIEVAL	

* *

May 21, 2009

Mail Stop Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPEAL BRIEF

Sir:

Appellant hereby appeals to the Board of Patent Appeals and Interferences from the last decision of the Examiner.

TABLE OF CONTENTS

(I) REAL PARTY IN INTEREST

The real party in interest is British Telecommunications public limited company, a corporation of the country of the United Kingdom.

(II) RELATED APPEALS AND INTERFERENCES

The appellant, the undersigned, and the assignee are not aware of any related appeals, interferences, or judicial proceedings (past or present), which will directly affect or be directly affected by or have a bearing on the Board's decision in this appeal.

(III) STATUS OF CLAIMS

Claims 1-10 are pending and have been rejected. No claims have been substantively allowed. All of rejected claims 1-10 are being appealed.

(IV) **STATUS OF AMENDMENTS**

A Response requesting reconsideration of the Final Rejection was filed on February 25, 2009. An Advisory Action was issued on March 24, 2009 continuing to reject all claims.

(V) SUMMARY OF CLAIMED SUBJECT MATTER

Each independent claim, each dependent claim argued separately, and each claim having means plus function language is summarized below including exemplary reference(s) to page and line number(s) of the specification.

1. A method for determining the semantic similarity of words in a plurality of words selected from a set of one or more documents, for use in the retrieval of information in an information system, comprising the steps of:

(i) for each word of said plurality of words:

(a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words [Fig.2, step 215; p. 6, ln. 20 to p. 7, ln. 18];

(b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word [Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

(c) generating a fuzzy set comprising, for word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b) [Fig.2, step 225; p. 8, ln. 5 to p. 9, ln. 15]; and

(ii) calculating and storing, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair [Fig.2, step 230; p. 9, ln. 16 to p. 10, ln. 20].

2. A method according to Claim 1, further comprising the step of:

(iii) adding a new document to said set of one or more documents and, using a set of words selected from said new document, performing an incremental update to said stored probabilities by means of steps (i) and (ii) performed in respect of said selected words using word sequences identified in said new document [p. 13, lns. 2-14].

3. An information retrieval apparatus for use in retrieving information from a set of one or more documents, comprising:

an input for receiving a search query [Fig. 1, 100, Fig. 2, step 200; p. 5, lns. 31-33];

generating means for generating a set of probabilities indicative of the semantic similarity of words selected from said set of one or more documents [Fig. 1, 105, Fig. 2, step 220; p. 7, ln. 19 to p. 8, ln. 4];

query enhancement means for modifying a received search query with reference, in use, to said generated set of probabilities [Fig. 1, 125, Fig. 2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

information retrieval means for searching said set of one or more documents for relevant information using a received search query modified by said query enhancement means [Fig. 1, 115, Fig. 2, p. 5, lns. 2-4],

wherein said generating means are arranged, in use:

(i) for each word selected from said set of one or more documents:

(a) to identify, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words [Fig. 2, step 215; p. 6, ln. 20 to p. 7, ln. 18];

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word [Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b) [Fig.2, step 225; p. 8, ln. 5 to p. 9, ln. 15]; and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair [Fig.2, step 230; p. 9, ln. 16 to p. 10, ln. 20].

4. An information retrieval apparatus according to Claim 3, wherein said query enhancement means are arranged to identify, with reference to said generated set of probabilities, a word having a similar meaning to a term of said received search query and to modify said search query using said identified word [Fig. 1, 125; p. 13, lns. 18-26].

5. An information retrieval apparatus according to Claim 3, further comprising updating means for adding a new document to said set of one or more documents and, using a set of words selected from said new document and word sequences identified in said new document, performing an incremental update to said generated set of probabilities in respect of words in said set of words [Fig. 1, 115; p. 13, lns. 2-14].

6. An information retrieval apparatus for use in retrieving information in an information system, comprising:

an input for receiving a search query [Fig. 1, 100, Fig. 2, step 200; p. 5, lns. 31-33];

generating means for generating a set of probabilities indicative of the semantic similarity of words selected from a sample set of one or more documents [Fig. 1, 105, Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4];

query enhancement means for modifying a received search query with reference, in use, to said generated set of probabilities [Fig. 1, 125, Fig. 2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

information retrieval means for searching said information set for relevant information using a received search query modified by said query enhancement means [Fig. 1, 115, Fig. 2, p. 5, lns. 2-4], wherein said generating means are arranged, in use:

(i) for each word selected from said sample set:

(a) to identify, in documents of said sample set, word sequences comprising the word and a predetermined number of other words [Fig.2, step 215; p. 6, ln. 20 to p. 7, ln. 18];

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word [Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b) [Fig.2, step 225; p. 8, ln. 5 to p. 9, ln. 15]; and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair [Fig.2, step 230; p. 9, ln. 16 to p. 10, ln. 20].

7. An information retrieval apparatus according to Claim 6, further comprising updating means for adding a new document to said sample set of one or more documents and, using a set of words selected from said new document and word sequences identified in said new document, performing an incremental update to said

generated set of probabilities in respect of words in said set of words [Fig. 1, 115; p. 13, lns. 2-14].

8. An information processing apparatus for use in an information processing apparatus, for use in an information system, for identifying information sets associated with a predetermined information category, the apparatus comprising:

generating means for generating, in the form of a matrix, a set of probabilities indicative of the semantic similarity of words selected from a sample set of one or more documents representative of the predetermined information category [Fig. 1, 105, Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4];

calculating means arranged to calculate, for each information set, a vector of values representing the relative frequency of occurrence, in the information set, of words represented in a matrix generated by the generating means [Fig. 1, 125, Fig. 2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

clustering means arranged to determine a measure of mutual similarity between pairs of information sets, using the respectively calculated vectors and the generated matrix, and to use the determined measures in a clustering algorithm to select one or more information sets to associate with the predetermined information category [Fig. 1, 120, Fig. 2, p. 5, lns. 1-2],

wherein said generating means are arranged, in use:

(i) for each word selected from said sample set:

(a) to identify, in documents of said sample set, word sequences comprising the word and a predetermined number of other words [Fig.2, step 215; p. 6, ln. 20 to p. 7, ln. 18];

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word [Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b) [Fig.2, step 225; p. 8, ln. 5 to p. 9, ln. 15]; and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair [Fig.2, step 230; p. 9, ln. 16 to p. 10, ln. 20].

9. An information processing apparatus according to Claim 8, wherein the clustering algorithm is a hierarchic agglomerative clustering algorithm [p. 12, lns. 21-31].

10. A method for determining the semantic similarity of words in a plurality of words selected from a set of one or more documents, for use in the retrieval of information in an information system, comprising the steps of:

(i) for each word of said plurality of words:

(a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words [Fig.2, step 215; p. 6, ln. 20 to p. 7, ln. 18];

(b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word [Fig.2, step 220; p. 7, ln. 19 to p. 8, ln. 4]; and

(c) generating, from the relative frequencies determined at step (b), a set of probabilities representative of the contexts in which the word occurs [Fig.2, step 225; p. 8, ln. 5 to p. 9, ln. 15]; and

(ii) calculating and storing, for each pair of words of said plurality of words, using respective probability sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair [Fig.2, step 230; p. 9, ln. 16 to p. 10, ln. 20].

(VI) GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. Claims 1-10 have been finally rejected under 35 U.S.C. § 102(e) as being anticipated over Choi.

(VII) ARGUMENT

A. Claims 1-10 which have been finally rejected under 35 U.S.C. § 102(e) are not anticipated by Choi

It is well established in the patent law that for a reference to anticipate a claim the reference must expressly or inherently disclose every limitation of the claim. *Rowe v. Dror*, 112 F.3d 473 (Fed.Cir. 1997). Since the cited Choi reference fails to teach (or even suggest) certain claim limitations of the present claims, the Examiner's rejection of the claims as being anticipated by Choi must be reversed.

Appellants' invention involves an information retrieval system that recognizes the semantic similarity of different words to determine whether one document has similar contents to another. For example, with no semantic knowledge or understanding a computer will not match the word "taxi" with the word "cab." In Appellants' invention it is possible for the system to determine the semantic similarity of words; the text of the document set is stemmed, optionally after the exclusion of the most and least common words from the document set, and the resulting word output is analyzed to determine a number of n-grams. Paragraphs [0055] - [0063] of the present application provide an example of how four sentences may be analyzed to generate a number of 3-grams (that is an n-gram for the case where n=3). Conversely, in Choi it is necessary for a human user to submit a set of keywords that have a semantic similarity in order that matches between different documents can be determined, for example to tell the system that "taxi" and "cab" have the same meaning.

The Final Office Action and the Advisory Action states that "[a]though the applicant's specification discloses the stemming algorithm, it is not mentioned in the claims." *See,* Final Office Action at page 3, Advisory Action at page 2. To the contrary each of independent claims 1, 3, 8, and 10 recite limitations directed to the stemming algorithm – which limitations are not found in the cited reference. For example, integers of claim 1 directed to the stemming algorithm include:

> (i) for each word of said plurality of words:
>
> (a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;
>
> (b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and
>
> (c) generating a fuzzy set comprising, for word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b) . . .

See, claim 1 (emphasis supplied). Moreover, as will be explained below, Choi does not teach or suggest these same claim limitations.

The Final Office Action states that paragraph [0002] of Choi inherently discloses limitation i(a) of independent claims 1, 3, 6, 8 and 10. *See,* Final Office Action at page 4.

> The present invention relates to a method of order-ranking document clusters using entropy data and Bayesian self-organizing feature maps(SOM), in which an accuracy of information retrieval is improved by adopting Bayesian SOM for performing a real-time document clustering for relevant documents in accordance with a degree of semantic similarity between entropy data extracted using entropy value and user profiles and query words given by a user, wherein the Bayesian SOM is a combination of Bayesian statistical technique and Kohonen network that is a type of an unsupervised learning.

See, Choi at paragraph [0002] (emphasis supplied). However, the cited portion of Choi does not disclose, for example the claim 1 limitation of "for each word of said plurality of words . . . indentifying, in documents . . . word sequences comprising the word and a predetermined number of other words." Choi merely discloses the conventional use of query words, and does not even mention identifying word sequences comprising the word and a predetermined number of other words.

The Final Office Action misapprehends Applicants' invention by asserting:

> "Finding word sequences" is anticipated by, "query words given by the user" (¶ 0002).

See, Final Office Action at page 4. This is a non-sequitur in that "query words given by the user" as used in the passage of Choi is not the same as "finding word sequences." Moreover, as noted above, the actual limitation of "identifying . . . word sequences comprising the word and a predetermined number of other words" is not found in the portion of Choi cited by the Examiner or anywhere else. Thus, Choi fails to teach or suggest limitation i(a) of the present claims.

The Final Office Action asserts that claim limitation i(b) of the present claims can be found at paragraph [0027] and Fig. 2, step 20, of Choi, stating that portion of Choi discloses "frequencies of keywords." *See,* Final Office Action at page 4.

> To accomplish the above objects of the present invention, there is provided a method of order-ranking document clusters using entropy data and Bayesian SOM, including a first step of recording a query word by a user; a second step of designing a user profile made up of keywords used for the most recent search and frequencies of the keywords, so as to reflect a user's preference; a third step of calculating entropy value between keywords of each web document and the query word and user profile; a fourth step

> of judging whether data for learning Kohonen neural network which is a type of unsupervised neural network model, is sufficient or not; a fifth step of ensuring the number of documents using a bootstrap algorithm, a type of statistical technique, if it is determined in the fourth step that the data for learning Kohonen neural network is not sufficient; a sixth step of determining prior information to be used as an initial value for each parameter of network through Bayesian learning, and determining an initial connection weight value of Bayesian SOM neural network model where the Kohonen neural network and Bayesian learning are coupled one another; and a seventh step of performing a real-time document clustering for relevant documents using the entropy value calculated in the third step and Bayesian SOM neural network model.

See, Choi at paragraph [0027] (emphasis supplied). However, the cited portion of Choi does not disclose the actual limitation of, for example, claim 1: "calculating a relative frequency of occurrence for <u>each distinct word sequence among word sequences containing the word</u>." Indeed, the cited portion of Choi discloses nothing about frequencies involving each distinct word sequence among word sequences containing the word, as required by claim 1, but merely <u>frequencies of the keywords</u>. Thus, Choi fails to teach or suggest limitation i(b).

A previous Office Action asserted that limitation i(c) could be found at paragraph [0143] of Choi.

> Clustering method includes k-nearest neighbor method, fuzzy method and the like. However, the present invention adopts a clustering method where documents are clustered by a statistical similarity, i.e., standardized distance between the two documents. In other words, a hierarchical document clustering where document cluster is formed through grouping documents having high statistical similarity, starting from each clusters made up of each documents expressed in terms of statistical similarity.

See, Office Action, dated November 28, 2007, at page 3. However, the cited portion of Choi does not disclose the actual limitation of, for example, claim 1: "generating a fuzzy set comprising, for word sequences containing the word, corresponding fuzzy membership values . . . " Indeed, once again, the cited portion of Choi makes no mention of word sequences at all. Thus, Choi also fails to teach or suggest limitation i(c) of the present claims.

As demonstrated above, the claim integers (i) (a) - (c) are not disclosed (or even suggested) by Choi and, thus, claim 1 and its dependent claims patentably define over Choi. Independent claims 3, 6, 8, and 10 contain similar claim integers and, therefore, these claims and their respective dependent claims also patentably define over Choi. Choi does not teach towards the solution provided by Appellants' invention; indeed Choi discloses the use of a manual method of providing a semantic link and thus Choi teaches ***away*** from the present invention.

In addition, with respect to claim 8, the Examiner has failed to identify where Choi discloses a clustering means or clustering algorithm. *See,* Office Action, dated November 28, 2007, at page 5. Accordingly, claim 8 patentably defines over Choi for this additional reason.

Dependent claims 2, 5, and 7 further require adding a new document to the document set and using a set of words selected from the new document to update the probabilities indicative of the similarity of words selected from the set (or sample) of documents. The Examiner has cited Choi at paragraphs [0166] and [0171] as disclosing this further claim limitation, however, the cited paragraphs have nothing to do with updating probabilities and therefore cannot anticipate these claims. *Id.*, at page 4.

Dependent claim 4 further requires that the query enhancement means identify a word having a similar meaning to a term of the received search query and then modify the search query using the identified word. The Examiner has cited Choi at paragraph [0087] as disclosing this further claim limitation, however, the cited paragraphs has nothing to do with modifying the search query based upon identifying a similar word. *Id.*, at pages 4-5.

Dependent claim 9 further requires that the clustering algorithm of claim 8 is a hierarchic agglomerative clustering algorithm. The Examiner has cited Choi at paragraph [0035] as disclosing this further claim limitation, however, the cited paragraph of Choi is merely a description of Figure 5A-5D and does not disclose any algorithm but merely states that these figures "illustrate concepts of hierarchical clustering for a statistical similarity between document clustering and query words according to the present invention." *Id.*, at page 6.

For all of the above additional reasons, dependent claims 2, 4, 5, 7, and 9 are not anticipated by Choi.

CONCLUSION

In conclusion it is believed that the application is in clear condition for allowance; therefore, early reversal of the Final Rejection and passage of the subject application to issue are earnestly solicited.

Respectfully submitted,

NIXON & VANDERHYE P.C.

By: ______________________
Chris Comuntzis
Reg. No. 31,097

CC:lmr
901 North Glebe Road, 11th Floor
Arlington, VA 22203-1808
Telephone: (703) 816-4000
Facsimile: (703) 816-4100

(VIII) CLAIMS APPENDIX

1. A method for determining the semantic similarity of words in a plurality of words selected from a set of one or more documents, for use in the retrieval of information in an information system, comprising the steps of:

(i) for each word of said plurality of words:

(a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;

(b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) generating a fuzzy set comprising, for word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) calculating and storing, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

2. A method according to Claim 1, further comprising the step of:

(iii) adding a new document to said set of one or more documents and, using a set of words selected from said new document, performing an incremental update to said stored probabilities by means of steps (i) and (ii) performed in respect of said selected words using word sequences identified in said new document.

3. An information retrieval apparatus for use in retrieving information from a set of one or more documents, comprising:

an input for receiving a search query;

generating means for generating a set of probabilities indicative of the semantic similarity of words selected from said set of one or more documents;

query enhancement means for modifying a received search query with reference, in use, to said generated set of probabilities; and

information retrieval means for searching said set of one or more documents for relevant information using a received search query modified by said query enhancement means,

wherein said generating means are arranged, in use:

(i) for each word selected from said set of one or more documents:

(a) to identify, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

4. An information retrieval apparatus according to Claim 3, wherein said query enhancement means are arranged to identify, with reference to said generated set of probabilities, a word having a similar meaning to a term of said received search query and to modify said search query using said identified word.

5. An information retrieval apparatus according to Claim 3, further comprising updating means for adding a new document to said set of one or more documents and, using a set of words selected from said new document and word sequences identified in said new document, performing an incremental update to said generated set of probabilities in respect of words in said set of words.

6. An information retrieval apparatus for use in retrieving information in an information system, comprising:

an input for receiving a search query;

generating means for generating a set of probabilities indicative of the semantic similarity of words selected from a sample set of one or more documents;

query enhancement means for modifying a received search query with reference, in use, to said generated set of probabilities; and

information retrieval means for searching said information set for relevant information using a received search query modified by said query enhancement means,

wherein said generating means are arranged, in use:

(i) for each word selected from said sample set:

(a) to identify, in documents of said sample set, word sequences comprising the word and a predetermined number of other words;

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

7. An information retrieval apparatus according to Claim 6, further comprising updating means for adding a new document to said sample set of one or more

documents and, using a set of words selected from said new document and word sequences identified in said new document, performing an incremental update to said generated set of probabilities in respect of words in said set of words.

8. An information processing apparatus for use in an information processing apparatus, for use in an information system, for identifying information sets associated with a predetermined information category, the apparatus comprising:

generating means for generating, in the form of a matrix, a set of probabilities indicative of the semantic similarity of words selected from a sample set of one or more documents representative of the predetermined information category;

calculating means arranged to calculate, for each information set, a vector of values representing the relative frequency of occurrence, in the information set, of words represented in a matrix generated by the generating means; and

clustering means arranged to determine a measure of mutual similarity between pairs of information sets, using the respectively calculated vectors and the generated matrix, and to use the determined measures in a clustering algorithm to select one or more information sets to associate with the predetermined information category,

wherein said generating means are arranged, in use:

(i) for each word selected from said sample set:

(a) to identify, in documents of said sample set, word sequences comprising the word and a predetermined number of other words;

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

9. An information processing apparatus according to Claim 8, wherein the clustering algorithm is a hierarchic agglomerative clustering algorithm.

10. A method for determining the semantic similarity of words in a plurality of words selected from a set of one or more documents, for use in the retrieval of information in an information system, comprising the steps of:

(i) for each word of said plurality of words:

(a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;

(b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) generating, from the relative frequencies determined at step (b), a set of probabilities representative of the contexts in which the word occurs; and

(ii) calculating and storing, for each pair of words of said plurality of words, using respective probability sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

(IX) EVIDENCE APPENDIX

None.

(X) RELATED PROCEEDINGS APPENDIX

None.